Hans Brügelmann (2005a)

International tests and comparisons in education performance: A pedagogical perspective on standards, core curricula, and the quality of schooling in the German education system[1]

[Ms. for: Rotte, R. (ed.) (2005): International perspectives on education policy. Nova Science Publ.: New York (forthcoming)]

The international comparative study PISA[2] has effected a profound change in the pedagogical world – at least one that will last for a period of several years. The problem can be described as follows: This comparison of basic skills in the concluding phase of compulsory schooling does provide important information for education policymaking, but the study itself – and its successor PIRLS[3] with reference to elementary schools – was overloaded with expectations that it cannot satisfy. Predictable disappointments will probably lead to a setback in empirical education research in general, much like the one during the 1970s. At that time, there were also evaluative studies, e.g., on compensatory support in preschool programs in the U.S. or on attempts to replace the tri-partite school system with comprehensive schools in Germany, that were unable to meet the expectations placed in them or to realize what they had promised.

Even though the authors of PISA have repeatedly emphasized how limited the significance of their findings is, public debate on the study has tended to overinterpret the findings. The following three limitations are readily overlooked:

- PISA does not specify how good or bad our schools actually are.

PISA illuminates *limited* aspects by means of certain *selected* instruments and allows us to take stock of *momentary* conditions. In this sense, the results are helpful as a warning to reflect on certain developments – but nothing more.

- PISA does not specify the reasons for the strengths and weaknesses of various schooling systems.

Correlations exhibited in PISA provide well-founded hypotheses on possible reasons and significant conditions behind the differences between and within educational systems. In this sense, the interpretations are helpful in the search for explanations, indicating points that could be subject to closer observation – but nothing more.

- PISA does not specify what is to be done to improve the quality of instruction in individual schooling systems.

The conclusions the authors draw are interesting interpretations that are supported and limited by the specific context of their particular discipline and position. In this sense, the recommendations are helpful as well-founded proposals to take steps in trying out certain suggestions – nothing more.

Considered in a positive light, PISA is important for the following two reasons:

---

[1] Translated into English by Thomas LaPresti. I wish to thank Georg Lind, who provides information on developments in education on a regular basis (Bildungs-Info, University of Constance) and through whose work I was able to use a number of not readily accessible sources. My thanks also go to Hans Werner Heymann for his conscientious critique of an earlier draft. Parts of this article were published in Brügelmann (1999) and (2003a+b); others expound on themes in Brügelmann (2004/05).

[2] Cf. OECD (2001; 2002).

[3] Cf. Martin et al. (2003).

- The comparative findings provide a substantial *heuristic* aid in locating problems and in searching for explanations of and solutions to such problems.

- The prominence of the experts and the political status of the project guarantee public interest in educational policy issues, a prerequisite for inducing change into such a static system as the school.

Still, the commendable study should not be credited with achievements it did not attain. The main problem has to do with the criteria for measuring the success of schools; more precisely, it is the lacking consensus on the definition of valid standards. Evaluating the quality or the effects of instruction requires normative guidelines.

In the context of the debate on educational policy in the last two or three years, two concepts have had far-reaching effects, esp. in Germany – 'core curricula' (cf. Böttcher/Hirsch 1999) and 'educational standards' (cf. Klieme et al. 2003). The reasons for the exceptional popularity of these concepts are related to the problems these instruments promise to solve. The promises include the following:

➔ disencumbering syllabi and focusing them on fundamental learning objectives;
➔ standardizing instruction in order to achieve more equity;
➔ improving instructional quality;
➔ guaranteeing a common basic education for all young people;
➔ ensuring minimum levels of student achievement in major subjects;
➔ evaluating achievement in the various subjects in an objective and more differentiated way.

These are noble goals which relate to problems in schools that are certainly to be taken seriously. This article will examine how plausible the promises of core curricula, standards, and testing programs to solve such problems actually are. To this end, I will make use of comparative empirical data from countries in which these concepts have already been introduced or have been practically implemented. In addition, I will refer to some past experiences with similar concepts that have vanished into oblivion at an astoundingly rapid rate (cf. Brügelmann 2003a).

At the outset, though, two general observations are in order here.

A number of experts demand a paradigm switch from controlling 'input' to controlling 'output'. According to this view, the main control instruments of educational policy in use throughout continental Europe, namely, the central guidelines for syllabi and lesson plans and the qualification requirements for teachers, have failed their purpose. Furthermore, there is a greater necessity to check the actual performance of students in individual subjects (output). This view maintains that Germany's relegation within the ranks resulting from PISA indicate that things look bad for 15-year-old German students even as far as basic skills are concerned.

At first glance, this line of argument seems reasonable. Yet, it ignores the fact that German fourth graders, with their fifth to twelfth places among 35 nations, achieved significantly better results in PIRSLtests of reading comprehension. These results were average for countries in the European Union, higher than the OECD average, significantly higher than for PISA and especially in comparison to earlier studies in 1970 and 1991. But elementary schools and secondary schools are 'regulated' in the same way. So how can regulating input be the reason for the poorer achievements of secondary school students in the international comparison?

The second observation is that the concepts used in the debate suggest a clarity that is nonexistent. What in the educational system actually is input or output in which sense and to which end, is an issue that is not sufficiently explicated (see Herrmann 2004). The simplified models overlook the numerous intermediary points and translating activities along the way

- from fundamental decisions on educational policy
- to guidelines, syllabi, and lesson plans,
- textbooks and workbooks,
- school programs and profiles,
- plans for organizing subject matter and concrete instruction units,
- the actual carrying out of instruction by individual teachers,
- individual activities of students,
- their immediate experience from such activities,
- up to short-term learning effects, and finally,
- to long-term learning success.[4]

Thus, the 'input' for the learning activity of students in a concrete sense is classroom instruction with its varied determinant factors.

Even such a short sketch of inconsistencies indicates that it is necessary to examine more closely which solutions core curricula and achievement standards provide for the problems in German schools – and whether the phenomena diagnosed as 'problems' are always actually to be interpreted as such.

Let us now consider the arguments in favor of core curricula and standards point by point.

1. Disencumbering syllabi and focusing them on fundamental learning objectives

A reduction of the long-lamented overabundance of content is to be achieved by introducing core curricula. A "core curriculum … defines knowledge that must be imparted to all children" (Böttcher 2002, 26). Here, it becomes readily apparent that even the proponents of the paradigm switch from controlling input to controlling output (op. cit., 17) only carry it out in a halfhearted way. The father of the core curriculum, American educationalist E. D. Hirsch (1997), even formulates his curriculum explicitly as a plan involving certain subject matter. Böttcher also speaks of a canon of content that involves methodological freedom and is "divided into subjects or groups of subject matter" (op. cit., 28). He demands greater "concreteness and clarity" (27) for this canon, which he wishes to make "mandatory" (30), but, on the other hand, he wishes to impose temporal restrictions on it (e.g., 60% of instruction time; 29).

In the U.S., syllabi actually are underdeveloped because the system is traditionally regulated at the local level and because the increasing attempts at central regulation in the last few decades have all been unambiguously output-oriented. In this sense, then, it is certainly reasonable to provide content specifications at the national level with a more pronounced profile. This context also explains the positive results (if not consistently so) that have been reported on the evaluation of first trial runs with core curriculum in the U.S. (cf. Becker 2002).

---

[4] On the widely diverse changes of the curriculum en route from the book author to the child, cf. my more detailed analysis following MacDonald/Walker (1976): Brügelmann (1978).

Experience with the 'national curriculum' in England, however, does not encourage making such guidelines mandatory at the national level. In his last annual report, even the director of the English evaluation agency OFSTED regretted a marked tendency to restrict instruction to mathematics and language, and consequently, to offer instruction in history, geography, art, and music[5] that he evaluated as being, for the most part, "flat and ordinary". He noted a significant decrease in student achievement in these subjects.

Focusing on the (apparent) essentials always involves the risk of restricting the curriculum, especially if performance in so-called main subjects is somehow associated with sanctions (cf. point 8 below). A survey undertaken in the U.S. showed that 80% of teachers increasingly spend more time on subjects that are tested and less time on other subjects.[6] This 'teaching to the test' not only affects the content of the curriculum, but also the style of teaching and learning.[7] Answers must fit into a pattern of 'correct' vs. 'incorrect', and content forfeits its inherent value.

Hameyer/Heckt (2004) correctly note that there are ideas and examples of concepts of a good, basic education from debates in educational theory in the 1950s and 60s[8] and from the debate on curriculum around 1970[9] available that are more perceptive than many of the hastily formulated current suggestions for standards. With reference to suggestions for language instruction from Baden-Württemberg, Wespel (2004) also criticizes the discrepancy[10] between standards and concrete 'sample problems' that often lead to a drilling of readily testable skills in easily assessable areas.

Educational standards that are really worthy of being called such must, then, encompass more than achievements in certain subjects. Above all, they must also determine criteria for the quality of educational *processes* in which students can actually appropriate what is demanded of them.[11]

2. Standardizing instruction in order to achieve more equity

Again and again, parents and parent organizations in Germany criticize regional differences in instruction. Interestingly enough, the emphasis is on differences *between* the 16 German states and much less on variances *within* the individual states.[12]

Especially with reference to PISA results, the high correlation between success at school and a specific social stratum, whose effects supposedly cumulate over several stages of a learner's biography, is a further object of criticism. At the outset, a substantial factor is the widely diverse stimulation provided by family milieus. It has also been demonstrated that the same (test) performance is assessed differently in the context of decisions relevant to selection for the tripartite secondary system.[13] This is doubly problematical since the learning ecology of the individual school or school type is especially important for learning success insofar as the

---

5 Cf. the report by Clare (2004).

6 Cf. Pedulla et al. (2003) and Stecher/Barron (2001).

7 On more detailed consequences of teaching to the test and of an overemphasis on basic skills, cf. Resnick/Resnick (1992, 37-75).

8 Cf., e.g., Klafki's argument in favor of content-transcending 'categories' instead of a material or formal education.

9 E.g., imparting key concepts and basic structures instead of isolated facts.

10 As an example, it is readily foreseeable that a standard assignment like writing a text for a picture story will narrow the perspective of a standard related to how children learn to control the writing process in an increasingly self-responsible way if teachers correctly assume that the quality of their instruction will be measured by the results of the assignment.

11 Cf. the formulation of demands children make on their education in Grundschulverband (2003) and 5 below.

12 Cf. the pronounced overlapping of distributions even in the comparison between high achieving Bavaria and low achieving Bremen in Bos et al. (2004, 61).

13 Cf. Bos et al. (2004, 191ff.).

stratum-specific ascriptions to various school types also lead to the provision of divergent learning opportunities :

> "Even with the same basic cognitive abilities and identical socio-economic status, the achievement of a student at a college-preparatory secondary school (Gymnasiasten) is 49 points higher than that of a student at a less prestigious secondary school (Hauptschüler)" (Baumert et al. 2001,182).

Still, it remains questionable whether stronger central regulation can actually control the decisive factors at the local level. The experiences gained with other educational systems can only make one feel skeptical on the issue.

After the political *Wende* in Germany in 1989, there was a rare opportunity to compare the academic achievement of students from two educational systems that differed greatly with respect to the intensity of central regulation. We undertook such a comparison on writing between the Federal Republic (West) and the GDR (East) and studied in exemplary fashion spelling in dictation and in independent texts at elementary school (grades 1 through 4). For dictations employing common basic vocabulary in the GDR sample, the error rate was actually lower than for the same dictations in the Federal Republic. But for the independent texts, the range of spelling achievement was about the same in both systems[14] – even though spelling instruction in the GDR had been prescribed and regulated in detail. The international study on reading skills conducted by the IEA in 1991 also established similar performance between 9-year-old students in the former GDR and their peer group in the Federal Republic.[15]

Delving a bit further back into the past, we find surprising evidence that even such a hierarchical and rigidly regulating system like that of the Nazi dictatorship was not able to exhaustively control the educational system – despite several attempts at *Gleichschaltung* in the twelve years between 1933 and 1945.[16]

In the English schooling system, a system that controls the observance of its highly structured 'national curriculum' by means of nationwide tests at regular intervals, student achievement varies more greatly than in Bremen, the German state that exhibited the highest variance in the IGLU-E tests of the PIRLS (cf. Bos et al. 2004, 63).

In addition, the PISA analyses show that, in general, centrally directed testing systems do not result in better achievement or in lesser variance of results (Baumert et al. 2000, 345-346).

In institutions that are so greatly dependent on personal relationships such as the educational or health care systems, standardization can only be achieved with a certain variation. To promote equal opportunities for students, positive discrimination in allocating resources can be applied by, e.g., granting special allowances to areas with considerable socio-economic problems. Thus, standardized tests are not needed for the specific support of schools with special problems. Sufficient for such decisions are social statistics on the economic conditions in the area served by a particular school.

---

[14] In the dictations, error rates varied less widely in the different grade levels of elementary school in the GDR (SD 11.1 – 19.6 vs. 14.7 – 26.6 for the Federal Republic); in independent texts rates were much the same (SD 5.6, 10.5, 13.5, 22.9 vs. 6.8, 8.6, 13.2, 23.4 in four elementary school grade levels), see Brügelmann/Richter (1994, 137).

[15] Cf. Lehmann et al. (1995, 216); in the 8th grades of the East German comprehensive school, variance was less than in the the tripartite West German system (op. cit., 219), but with a view to the findings in the elementary school, this effect is probably due to the differences in school organization than in the systems of control.

[16] On elementary school, cf. Götz (1987).

3. Guaranteeing a common basic education for all young people

Cultural competence is a high ideal, especially in a pluralistic society. With his concept of general education, Heymann (2003) explores the ability of subcultures and of different generations to engage in dialog with one another from several perspectives.

But can a core curriculum that is defined in terms of subject knowledge[17], that prescribes certain historical events, poems, and dramas as subject matter for instruction in secondary schools and the study of individual plants and animals in elementary schools, can such a content-oriented canon guarantee a common basic education?

The proponents of a canon neglect the fact that the same content can *mean different things* to different students with their specific background experience. Conversely, they also ignore the fact that comparable experience can be gained from widely varied content. In literature class, why should 17-year-old students not be allowed to choose which love poem (from the Baroque period or otherwise) they would like to study, as long as they are required to discuss their reasons and their reading experience? The same applies to subject matter in elementary schools. Thus, instead of concrete content, categories or criteria that can be conceived of abstractly in various forms should be established. The main issue is a higher quality of learning *opportunities*, i.e., the emphasis should be on the educational demands and needs of the children, not so much on the qualification requirements of society (cf. Grundschulverband 2003).

To allow for more openness in instruction, content, e.g., could be made more accessible to an individualization 'from below', from the students, instead of demanding too much of teachers by requiring a differentiation 'from above'. Features in common would then be established by the exchange among the students, not by imposition from the teacher. There is no (longer) a unified or predominant culture. The important thing is that the various subcultures maintain the ability to engage in dialog with one another. Thus, the school is not to be thought of as a place for giving lectures or even for making converts, but, rather, as a forum for the encounter and interaction of various cultures and generations. It institutionalizes this mutual exchange as a means of learning with and rom one another. This also applies to the appropriation of social conventions.[18]

In addition, the more concrete and detailed content is specified for guaranteeing common ground, the more acute the problem of transfer to other content and other situations becomes. This has become apparent in exemplary fashion by the current debate on empirical studies of the educational value of Latin.[19]

4. Ensuring minimum levels of student achievement in major subjects

> "Educational standards specify which *comopetencies* children or youths should have attained by a certain grade level." (Klieme et al. 2003, 4).

Specifying the same requirements for all students at certain defined points in time is intended to prevent individual students from slipping through the educational net and failing at the labor market due to insufficient qualifications. But, what are 'minimum levels of achievement', how can they be defined?

---

17 See Hirsch's suggestions, referred to in Böttcher's arguments.
18 Cf. the didactic conception of Gallin/Ruf (1998).
19 Cf. the controversy between Haag/Stern (2000), Strunz (2003), and Westphalen (20003). See 5 below.

In the U.S., the Bush administration is attempting to reach this goal with a law known by the promising name of 'No Child Left Behind'. First analyses have shown that in Minnesota, a state that had results much higher than average in nationwide achievement tests of recent years, over 80% of the schools have not attained the levels specified by the state.[20] Since achievement cannot be increased as simply as the education policymakers had promised, other states have already lowered the requirements in order to avoid sanctions imposed by the 'No Child Left Behind' law of the federal government.[21]

In Germany, PISA and PIRLS have clearly demonstrated the problem of inadequate student achievement in reading, to name one example.[22] But can the adoption of systematic elements from other countries, as so often propagated, really provide a solution to this problem? In the PIRLS study, Bremen, for instance, had the worst results of the 16 German states, primarily because of especially poor performance at the lower levels. Yet, on the international scale, the lowest rank (5th place in the percentage ranking) corresponds to achievement levels attained in countries such as New Zealand (internationally, the model system for promoting reading[23]) and England (3rd place in PIRLS) and the U.S. (9th place), although in the latter two countries there have been standards and achievement tests at regular intervals at the national level for the past 10 to 30 years.

Achievement variance is a phenomenon that can be observed in all schooling systems (see 2 above). Our study of reading in North Rhine Westphalia, LUST[24], demonstrates that from grades 2 to 4 the 10% low achievers of the same grade lag behind the 10% high achievers by three to four grade levels. Data from PIRLS demonstrate that the variance in other countries is even considerably higher than in German elementary schools.[25] Studies on first graders have shown that this wide range of achievement is already accounted for by the children's various preschool experiences.[26] In the light of such heterogeneousness, how can reasonable threshold values be established for the end of second or fourth grade without demanding much too little of the one group and ostracizing the other group at the outset?[27]

And a further problem arises. No one can actually say at which level the necessary *basic* qualifications are to be defined. In a pilot study, we undertook a test of basic reading abilities in a trade agency. According to the results of that survey, almost 90% of the fourth-grade students could already compete with the basic reading ability of craftspeople attending master courses, even if their 5% lowest achievers were neglected.

Thus, it is not at all surprising that expert opinions on establishing a workable basis for reading competency at the end of elementary school or of compulsory schooling are widely divergent. The International Adult Literacy Study of the OECD (1995) comes to the conclusion that people with very different levels of objectively tested reading ability get along well professionally and in everyday life. This study and additional ones[28] also demonstrate that the process of learning how to read is not terminated with the first grade, nor with elementary school, nor with the end of a person's formal schooling at all. Thus, even for the most basic

---

20 See ➔ http://www.twincities.com/mld/pioneerpress/living/education/8041554.htm [27.2.2004]

21 See, e.g., the report from the New York Times on May 22, 2003 on the reduction of requirements in Texas http://www.nytimes.com/2003/05/22/education/22EDUC.html [24.5.2003] and the report from CNN on October 12, 2003 http://www.cnn.com/2003/EDUCATION/12/10/states.education.ap/index.html [12.12.2003]; see also Ratzki (2003).

22 Elsewhere, I have dealt with the issue that this is not to be equated with a decline in scholastic achievement, but, instead, is a consequence of considerably increased requirements for use of written language in everyday and professional life (cf. Brügelmann 1999, 10-20, 120-126).

23 Cf. Rossa/Rossa (1995).

24 Brügelmann, H. (2003a).

25 PIRLS showed variances for achievements of the average two thirds at the end of 4th grade covering around two grade levels; for the highest 5% vs. the low 5%, this covered 5-6 grade levels. Cf. Bos. et al. (2004, 62-65).

26 Brügelmann (1983, 201), Rabenstein et al. (1989), and Richter (1992).

27 Critique on the 'No Child Left Behind' law is in a similar vein; cf. Marshak (2003).

28 Cf. Brügelmann (2004).

skills it seems questionable whether it is reasonable and even possible to describe the termination of learning processes by means of standards.

Nor has the question of whether there are threshold values for workable basics without which a secondary school student cannot proceed to learn successfully actually been resolved. Learning is a cumulative process, but not a linear one. Test scores in a certain subject are inadequate factors for predicting someone's future abilities since these are also dependent upon personal development.[29]

Vor allem aber ist die Annahme naiv, eine Anhebung des Leistungsniveaus im unteren Bereich könnte die Jugendarbeitslosigkeit beseitigen, indem sie zu einer Erhöhung der Einstellungen führt. Sie mag die relativen Chancen des Einen auf Kosten der Anderen verbessern, aber die ständigen Entlassungen von Arbeitnehmern sind selten die Folge von Kompetenzschwächen, sie sind die Folge des Rationalisierungsdrucks in der Wirtschaft. Je mehr SchulabsolventInnen wissen und können, umso höher wird das Niveau der sog. „Mindestleistungen" angesetzt werden. Rechtschreibtests lassen sich auch so konstruieren, dass sie noch unter den leistungsstärksten 10% „VersagerInnen" selegieren.

Wir brauchen deshalb empirische Studien, die alltagsnah bestimmen, welche schriftsprachlichen Fähigkeiten in verschiedenen Lebensbereichen erforderlich sind –und wie weit diese bis zum Abschluss der Pflichtschule entwickelt sein müssen (oder auch noch „on the job" erworben werden können).

5. Improving instructional quality

American states that made awarding diplomas, supplementary pay for teachers, and the allocation of finances to schools contingent on student results in annual tests in the 1990s have repeatedly reported an increase in achievement. At the beginning of 2000, 28 states – more than half of the United States – had committed themselves to this program. Yet, a study undertaken by Amrein and Berliner (2002) demonstrates that the increase reported by the states usually only applies to the limited area of those tests that are already established practice in the particular state. On independent tests like those testing aptitude for college-level study, there was a *decrease* in test scores in 2/3 of the 28 states. Increasing dropout rates were also reported. In other words, low achievers had completely dropped out of the system, additionally driving test scores upward, but without any actual improvement in 'output'.[30]

Because of the varied conditions in individual classrooms, it would be necessary to gauge students' learning progress from their own individual needs in order to adequately judge the quality of instruction.[31] Socio-economic factors relative to the school district, like those amassed for PISA and PIRLS, are too superficial. A simple counterexample that illustrates this point would be the heterogeneous compositions of parallel classes in the same school, in which one class consists of children from a neighborhood with buildings made from prefabricated slabs, the other of children from a neighborhood where the parents own their homes.

---

[29] For more detail, see Brügelmann (2003d).

[30] Cf. the summary and commentary by Winter in the New York Times on December 28, 2002 and in more detail 8 below.

[31] On problems related to the concept of learning 'profit', cf. Kupermintz (2003). He refers, above all, to the questionable assumption that even under the same subject-related preconditions learning progress is supposedly solely dependent on the teacher and that advances are additive and can be modeled or calculated in a linear way.

Even more fundamentally, it is questionable whether instructional quality can actually be controlled by measures from a centralized authority. The concept of 'control' itself is associated with power fantasies[32] that (should they be acceptable at all) are certainly not realistic in the light of the education reforms of the last thirty years. In this sense and despite all lip service paid to notions of autonomy, as innovation strategies or as sociological concepts of organizations centralized control models are, in the final analysis, naïve.[33] In a poll of 214 younger secondary school teachers in Baden-Württemberg on the induction of educational standards, about 70% agreed with the statement that "my classroom instruction will not be profoundly affected" (Rauin 2003).

Centralized tests are overloaded with expectations, while opportunities for self-evaluation are simply overlooked. Evaluations of systems are important, but in a more differentiated composite model[34] more emphasis would have to be placed on an evaluation closely related to classroom instruction, an evaluation that in addition to effects would also gauge the quality of processes.[35] In addition to more general competency levels that only become concretized directly at schools, the new Finnish educational standards also define requirements for the learning environment and for the style of classroom instruction. The arguments behind such requirements are of a normative nature (related to principles of human interaction) and not simply instrumental (with a view to results).

Such standards for high quality instruction can be put into concrete terms as criteria of organizational form and of the material surroundings, of mutual consent on rules and of the quality of stimuli provided by teachers. These should be characterized in the following ways:

- they should stimulate children to do self-assigned and self-responsible work, e.g., by means of joint planning sessions in the mornings or by individual learning 'contracts';
- they should challenge familiar ways of thinking and commonly available patterns of behavior, e.g., by confronting the students with other perspectives in group discussions or through repeated queries from the teacher at the workplace;
- they should facilitate widely diverse activities, e.g., by providing opportunities to realize individual plans or by offering prepared elective activities;
- they should demand and support an exchange of individual experiences and results, e.g., with institutionalized forms of reports;
- they should promote reflection on one's own work, e.g., by means of self-evaluation at regular intervals and constructive criticism of the work of others;
- they should ensure helpful support when difficulties arise, e.g., by a tutoring system and by regulated access to the teacher;
- they should substantiate rules and directions with reference to their function (avoiding disturbances of others or impediments to work), not in a authoritarian manner.

Finally, greater emphasis on orientation to processes is also important for educational research, for it can only develop options for improving instruction by studying processes more closely.

---

[32] Cf. Herrmann (2003, 43).

[33] This was maintained by the German Council on Education as early as 30 years ago. Deutscher Bildungsrat (1974).

[34] See 9 below.

[35] Cf., on the one hand, „Aufruf für einen Verbund reformpädagogisch engagierter Schulen" on more substantial self-evaluation ➔ www.BlickUeberDenZaun.de and, on the other, the *process* standards for *opportunity-to-learn* of NCTM in the U.S. (ref. Klieme et al. 2003, 24, 25, 28-30 – highly praised here, even though – with no further justification – these scholars opted for an output-orientation of standards).

Low cost instruments that are robust and can be used in everyday schooling without undue complications are to be developed in order to put into practice the demand for externally supported self-evaluation. These instruments are to be complemented by comparative data from representative samples that are differentiated according to subgroups and milieu conditions so that teachers are provided with concrete data for calibrating their assessment criteria.[36] The significance of this approach will become evident in the next section.

6. Evaluating achievement in school subjects in a more objective and differentiated way

As instruments for assessing achievement, grading has been controversial for a long time.[37] There is, for instance, only a limited correspondence in ranking between grades and standardized test scores. For PIRLS, Bos et al. found a correlation in reading comprehension of only .56**. One consequence of the PISA study was a widespread rebuke of teachers. First of all, teachers[38] were said to be unable to recognize low reading achievers, and secondly, the grades they assigned seemed to show an unacceptably wide variance in relation to the same level of achievement as measured in standardized tests.[39] But the findings reported in this context can be explained in a variety of ways.

1. Teachers are inept at making diagnoses; they cannot differentiate between different levels of achievement.
2. Teachers are diagnostically competent, but their criteria are so varied that the same achievement level is judged differently in different classes.
3. Tests and teacher evaluations assess different things[40]; there is only a partial correspondence between the particular basis for evaluation and the assessments themselves.

Data resulting from our reading survey LUST can help to check the first two hypotheses.[41] The first line in table 1 demonstrates the correlations between reading achievements in a reading test and the classroom grade for reading in the entire sample. They are at the level reported by PIRLS. If these values are compiled for each class and class averages are correlated (line 2), then the correspondence is reduced considerably. If, on the other hand, the correlations for each class are analyzed separately and only then is the average of these correlations determined (line 3), the correspondence is significantly higher.

Table 1: Correlations of reading test results with classroom grades in reading	3rd grade	4th grade
individual correlations across all classes[42]	.63**	.56**
class averages: classroom reading grade and test score[43]	.32**	.17**
internal class correlations: reading grade and test score[44]	.70**	

[36] On the basis of a proposal by Wilfried Metze (Berlin), we developed this procedure with reference to reading and successfully tested it as an aid to self-evaluation with more than a thousand teachers in North Rhine Westphalia. Cf. Brügelmann (2003c) and ➔ www.uni-siegen.de/~agprim/lust . A similar approach is applied in the testing program VERA, currently under development by Helmke et al. for several German states (see Schwarz 2003).

[37] Cf. Ingenkamp (1971!).

[38] At least in secondary schools (Baumert et al. 2001, 119-120).

[39] Cf. Baumert et al. (2003, 70-72).

[40] On varying results of tests in mathematics, cf. Ratzka (2004).

[41] In a similar vein: Schrader/Helmke (2001, 50) and Köller (2002).

[42] Per grade, ca. 6,000 children; results in PIRLS: .56**, see Bos et al. (2004, 204).

[43] In each case, 250-300 classes.

[44] Partial sample: 140 classes.

These findings support the second hypothesis. Most elementary school teachers do seem to be capable of recognizing distinct reading achievement levels[45], yet, they assign grades with reference to class-related criteria. From a pedagogical perspective, any other action would be hardly conceivable since the achievement is so widely variant between classes.[46] On the other hand, marks – especially at those transitional positions in German schools, the 4th, 9th, and 13th grades – are frames of reference for selective decisions that go far beyond the current class and school situation. The significance of such decisions accounts for the attractiveness of methods that promise comparable standards and objective data. But before teacher evaluation is simply replaced by tests, the function and value of such tests should also be scrutinized.

Here, Ratzka (2004) discovered that, with three different mathematics tests administered in elementary schools, only 41% of the students were assigned to the same group for all three tests and that also with the same type of task (word problem) on different tests varied results occurred –even on the same test (TIMSS for elementary school), depending on whether the problems had to be solved within a rigorous time limit.

In addition, even reputable tests display an astonishing number of mistakes, as was demonstrated in an analysis of mistakes in nearly 100 testing programs in the U.S. undertaken by Rhoades/Madaus (2003). In the meantime, American courtrooms have seen a number of trials in which decisions made on the basis of tests have successfully been disputed. It is generally ignored that test values do allow for relatively reliable assumptions about larger samples, but that, at the same time, their significance for individual cases can be extremely problematic due to measurement errors.

In this sense, then, it is justified to ask whether a single test is more valid than continuous observation by the teacher. Both methods are susceptible to malfunction, although in differing ways.[47] Thus, there are good reasons to employ tests as aids to repeatedly help calibrate a teacher's judgment. Standards involving competence levels and frequency distributions of achievement at certain fixed dates (e.g., at the end of the 2nd, 4th, and 8th grades) can help teachers to classify the development of individual students and of complete classes with reference to larger groups. If these data are divided into individual competence profiles, this will also facilitate a more detailed description of strengths and weaknesses in the individual achievement profile – in contrast to grades, which do not distinguish among achievements, but only indicate ranks of global achievement (in a certain subject).

Such competence-level models (in empirically validated form thus far only available for learning written language in elementary school[48]) can function as a general didactic orientation for recognizing early forms of achievement not yet conforming to norms and for developing a perspective on "the next step".
But as a scheme for assessing individual achievement, they are not without problems since they assume:
(a) that achievement is homogeneous (e.g., ability to apply a spelling rule, or lack thereof),
(b) that development is linear (one-dimensionality of only quantitative 'growth' with no roundabout routes),
(c) that development is always continuous (progression without regression),
(d) that achievement is not dependent on content and context,
all of which assumptions are rather improbable[49]. Thus, there is a risk that competence levels are not viewed as a heuristic instrument, but are used in an objectified fashion similar to the earlier use of IQ tests in special education.

---

45 Contrary to the PISA results (see above), this also applies to recognizing low achievers. See Brügelmann (2003c, 41).

46 In the LUST study, for example, the lowest 4th grade class had worse results than the best 2nd grade class (Brügelmann 2003c, 37). Such differences are not simply consequences of variably good instruction, but are also related to the composition of the classes, as already manifest in the different preconditions at the very beginning of schooling (cf. Rabenstein et al. 1989).

47 On the susceptibility of teacher assessments to malfunction, cf. the summary in Zielinski (1974).

48 Cf., e.g., the overview in Brinkmann (1997, chapter 2.2).

Above all, there is a risk that within the context of the debate on standards, individual levels are prescribed as norms and defined as achievement requirements for fixed dates.[50] In contrast, standards should be described as perspectives of development with different starting points from which demands can be made on each individual student and at the same time offering as much support as possible to students.

We should consider whether it might not be sufficient to *descriptively* indicate the competence level reached in report cards and to classify this as part of a representative distribution so that others can also assess the state of development. If the levels attained are also related to the particular learning requirements of the individual child and to the conditions in his or her surroundings, then the observer has a more distinct image of a student's *achievement* than with traditional grades or with a number of points on a comparative test related to specified achievement levels as standards.

7. A tentative appraisal

To summarize the major objections so far:

- The demand to establish specific content as *core curricula* overlooks the fact that young people can learn different things from the same content and the same things from different content.

- Regulation of *output* by focusing on narrowly defined learning objectives ignores the fact that the value of experience is dependent upon the quality of the processes by which that experience has been gained. Without standards for teachers' roles, for conditions of instruction, and for forms of social interaction and cooperation at school, education degenerates into an externally controlled form of training.

- To define *educational standards* (or to put them into practice) in such a way that the same achievement level is demanded of all students at the same time is illusory. Experiences related to subject knowledge already differ by several years of development when children begin their schooling, and this gap is not bridged during the years at school because all of the children make progress (the "caravan effect"). Assessment must be oriented to individual progress in development.

In addition, empirical findings provide little reason for optimism concerning the effects of the propagated change in the system to output-oriented control. And yet, we have not even begun to consider an especially significant problem: What is actually supposed to happen if the minimum requirements of established standards are not met by an educational system, by a school, by individual teachers or students?

8. Fortifying standards by imposing sanctions? Neglected lessons from history

The group of experts led by Eckart Klieme and charged with presenting a well-founded opinion to the German conference of education ministers has correctly maintained that standards of achievement are an instrument of educational policy for examining the performance of an *educational system* as a *whole*. But combined with an external assessment of students or of schools, they can have considerable undesirable side

---

[49] On developments related to spelling, cf. Brinkmann's findings (1997; 2003).

[50] The figures on the heterogeneousness of achievement for the same grade levels listed above (4) cast a critical light on such plans.

effects.[51] For example, it would be counterproductive to link failure to meet the requirements with sanctions imposed on individual institutions or persons ('high-stakes accountability').

First, a look at past developments:

- As early as 150 years ago, at a time of limited public funds (Crimean War!) students in Victorian England only received state subsidies for certified achievements in school subjects. Due to the corresponding definition of the requirements, this system of 'payment by results' led to significant savings in the national budget. But it seems not to have been a pedagogic success, for, otherwise, it certainly would have lasted much longer.[52]

- Around 1970, it was the advocates of programmed instruction and of an approach oriented to specific learning objectives who believed that predetermined learning effects could be guaranteed by "teacher proof" curricula that were described in detail. Was it only the obdurateness of the critics[53] that led to the situation that only a few years later hardly anything could still be heard of such promises?

- Over 30 years ago, the U.S. started a program of regular 'National Assessment of Educational Progress' in the major school subjects. In a parallel move, individual states also instituted statewide testing programs. Data from NAEP show that scholastic achievement, e.g., in reading, has hardly changed since 1970. For 17-year-old students, reading scores have only 'increased' from 285 to 288 points since 1971.[54] As early as in the 1970s, states like New York and California began to relate the performance of individual schools to expected values that were estimated from the learning dispositions of students and from socio-economic factors of the areas served by the schools so that unexpected low scores were an inducement to undertake studies and measures aiming at specific schools (Shepard 1979). At the time, more than 50% of the school districts in California claimed to have changed their programs on the basis of the data accumulated (Carlson 1979). More than 25 years later, the problems with fundamental performance still continue to evoke similar measures.

- Of the 115 U.S. school districts that decided to pay their teachers in accord with student achievement in 1978, more than half (68) had already abandoned this 'merit pay plan' five years later.[55]

- Also in the 1970s, companies in the U.S. made offers to individual school districts to increase student achievement to predetermined values for a fee. Since the endeavors met with little success, the school districts soon dispensed with this sort of 'performance contracting' (House 1975, 73-74). From the middle to the end of the 1990s, private companies like 'Edison' undertook another attempt. The tentative results after seven years are as follows (O'Reilly 2002):
  - in contrast to the promise of substantial increases, average student performance has not improved;
  - unsatisfied partners have terminated contracts ahead of time;
  - the company is not making profits, instead, its losses are constantly becoming more substantial;
  - after reaching their peak of $37, stocks have now fallen to a mere $2.

---

[51] Cf. a fundamental critique in Popham (2001; 2002). In the 1970s, Popham was a reputable developer of tests oriented to narrow learning objectives. Later, he became one of their most severe critics.

[52] Cf. House (1975, 73); see also Ascher (1996) and Aldrich (2000).

[53] Cf., e.g., the explicit opposing standpoint in the recommendations of the German Council on Education on promoting practice-oriented development of curriculum. Deutscher Bildungsrat (1974).

[54] Cf. Grigg (2003).

[55] Cf. Terhart (1997, 73) and references there to empirical data collected by Mumane/Cohen (1986) and Jacobsen (1989).

- Around the beginning of 2003, the English Department of Education and Skills was forced to admit that the increase in performance promised with the introduction of the national curriculum and nationwide tests had not been realized. After initial improvements in test performance, during the preceding three years only 65% of the students had managed to reach the defined threshold value ("The Guardian", January 5, 2003). Totally forgotten is the fact that more than 20 years ago an 'Assessment of Performance Unit' had already been established[56] to collect "firmer evidence about the standards of achievement nationally" (CERI 1978, 12). Apparently, the unit had no substantial effect on enhancing the quality of instruction, as the apparently still necessary current efforts demonstrate.

More recent studies in the U.S. provide evidence that scholastic achievements in states imposing more severe sanctions on institutions or groups of persons are worse than in states that impose lighter sanctions. 60-70% of the states with lighter sanctions are above the national average, but only 10-20% of the states with severe sanctions.[57] And where performance on standardized tests increased, it decreased for the most part on independent tests.[58] Linn (2000) was even able to show that performance only 'increased' as long as the tests remained unchanged. With the introduction of a new version of the same test, achievement was reduced to the original level. Afterwards, it 'increased' again, but only until the original form of the test was implemented again and performance once again 'decreased' to the original level.[59] These findings clearly demonstrate how punitive measures can encourage 'testing to the test', which can then interfere with or even jeopardize the desired content-related improvement of classroom instruction.

In other regions, especially the low achievers, for whose support the induction of standardized tests was actually intended, are at a disadvantage because in various ways teachers or schools have removed them from the test sample: In combination with inappropriate sanctions, standards and evaluations result in

- an enormous increase in the dropout rate,
- an increase in repeated grade levels,
- and increasing difficulties for economically and socially disadvantaged students to become admitted to schools, students who are then relegated to special schools.[60]

A further study differentiates between effects of 'high-stakes' testing on *students* (negative effects in general, especially for socially disadvantaged students) and on *schools* (increase in test performance, but combined with more pronounced selection).[61]

In a national survey of U.S. teachers on their perception of 'high- vs. low-stakes testing' conducted by Pedulla et al. (2003, 5-9), 8 of 10 teachers reported that they spent increasingly more classroom time on subjects tested and ever less time on subjects not covered by the tests. In addition, the tests led to

- an increase in students not promoted to the next grade level (according to 20% of teachers in states with high-stakes programs vs. 5% in those with low-stakes programs) and to
- an increase in dropouts (according to 25% of teachers in states with high-stakes programs vs. 10% in those with low-stakes programs).

In general, the teachers maintained that

- the time and energy needed was not worth the returns (about 75%),

---

56 Cf. Simon (1979); Kay (1979).
57 Cf. Sacks (1999), 89-90).
58 Cf. Amrein/Berliner (2002). Generally critical of the (side) effects of high-stakes testing: AERA (2003).
59 Cf. Linn (2000).
60 Cf. Darling-Hammond, L. (2003), summarized in Bildungs-Info (Georg Lind, University of Constance) on February 14, 2004; also to be found there is a reference to findings on increasing dropout rates in Clarke et al. (2000).
61 Cf. Schiller/Muller (2003).

- tests did not do justice to the performance of low achievers (90%), and
- colleagues on the teaching staff were able to attain higher test results without improving their classroom instruction (40%).

Finally, negative side effects that can result from exposing the weaknesses of those who do not reach the targeted scores must not be neglected.[62] A study in New Zealand demonstrated explicit negative effects of teachers' orientation to competition among their colleagues on the quality of classroom learning.[63] From several Anglo-Saxon countries there have been reports on teachers' attempts to circumvent the testing procedures[64] – even to the point of attempted fraud.[65]

9. Development of a differentiated evaluation system with discrete levels and functions[66]

This critique of the proposed solutions is not intended to deny the existence of the problems that they attempt to solve. What might be a feasible alternative? The prevailing idea of system monitoring (in itself, certainly necessary) is overloaded with expectations it cannot satisfy. The quality of classroom instruction is improved on location, and evaluation must be sensitive to the specific context of that instruction.

It seems most likely that a combination of various levels of self-evaluation and accountability to others can make the greatest use of the specific potentials involved in:

- **internal** and **external** evaluation
(familiarity with the situation and lack of anxiety vs. impartiality of judgment and necessary social pressures)

- **informal** and **formalized** methods
(adaptability and facility in completing surveys vs. objectivity and comparability of the data).

At the same time, such a combination of varying forms of evaluation will most likely be able to prevent extensive negative effects of the specific weaknesses and risks involved in the individual forms.

Here, we should keep in mind that concrete evaluative measures should aim at problems the solutions of which actually lie within the responsibilities of the institution or person being evaluated and that at the specific level examined there are possible courses of action available to improve the situation (at least the possibility of receiving support from external sources).

How, then, can the responsibility of the persons involved be more clearly perceived; what can these people do in a concrete sense to check the quality of their work and to improve it?

The following seven perspectives seem especially significant; they are readily feasible and can be developed from existing activities and then, step by step, combined to form an effective system of quality control:

---

62 Cf. the vivid description of the experiences of a German teacher who observed everyday schooling in the U.S. for one year in Leßmann (2003).

63 Cf. Ladd/Fiske (2003).

64 For example, in individual schools, school meals were changed to include higher glucose levels at the end of the school year, because it had been shown that higher levels can increase test scores – certainly a simpler method than improving classroom instruction. Cf. Figlio/Winicki (2003).

65 Kohn, A. (2002), cited in: Bildungs-Info (Georg Lind, University of Constance) on February 14, 2004.

66 The following suggestions have been taken from Brügelmann (1999). They have also been implemented in the German elementary school association's concept of a system of quality control. Cf. Bartnitzky et al. (1999).

- **Students** should make efforts to achieve a clear view of their own goals and of the successfulness of their own work, e.g., by committing themselves to certain goals and by self-evaluations of assignments, but also by regular reports on their own learning progress and on imminent goals.

- **Teachers** should observe and evaluate **learning processes** with a view to individual students and to the class as a whole; e.g., they should maintain a card file in which they record observations and sample assignments relevant to the development of the students and in which they continuously document learning processes so that they can discuss perceptions and explanations of progress and difficulties on an individual basis with the students and negotiate joint goals of future work with them.

- **Teachers** should also observe their own classroom instruction in order to apprehend and check the effects of their own work, e.g., to clarify their own demands and check their realization
  - in a first step, by means of a 'conversation' with oneself and by self-observation (e.g., in the form of a diary),
  - then, by requesting an 'external' perspective (that of a colleague, a trainee, a prospective teacher, or of parents).

- The **teaching staff** should develop and modify the school's program with a view to the specific needs of the **school**; the **school management** should stimulate, promote, and ensure the corresponding activities, by, e.g., periodically inquiring[67] people involved with the school:
  - "What would you identify as the specific strengths and weaknesses of our school (in a certain area)?"
  - "Do you have concrete suggestions for change?"
  - "Which procedures can we agree on?"

  External visitors should be asked the following:
  - "Do you notice anything particular about our school?"
  - "Are the goals and underlying assumptions of our school program convincing?"
  - "Where does our work lag behind the demands we make on ourselves?"

- The **school administration authority** can reinforce the external perspective with reference to **schools**, can acknowledge developments, provide support, formulate demands, by, e.g., visiting schools and asking the school management, teacher teams, and individual teachers questions like the following:
  - "What are your own goals?"
  - "How far are you able to realize these aspirations?"
  - "Which obstacles do you encounter in doing so?"
  - "Which goals have remained neglected?"
  - "Which realistic next steps are possible?"

- The **Department of Education** can identify general problematic areas and check political priorities with a view to the **entire system**, e.g., accumulate and present information needed to ensure material preconditions or to ensure that legal requirements are met. Such information could relate to secondary analyses of administrative statistics and to periodic studies of achievement and development status in anonymous samples of students along with surveys of the prerequisites, the process characteristics, and the basic conditions of classroom instruction.

---

[67] Cf., e.g., the instruments for self-evaluation developed by the Bertelsmann-Stiftung ➔ http://www.bertelsmann-stiftung.de/de/4056_5669.jsp [10.3.2004]; see also Schratz, M., et al. (2000): Qualitätsentwicklung. Vorschläge, Methoden, Instrumente. Beltz: Weinheim.

Thus, nationwide tests are to be conceived of as *one* building block in a much more comprehensive system. Their function is clearly to be understood as system monitoring and providing referential data for evaluation on location. But this form of stocktaking need not take place every year or even every second or third year. The period can be extended into the long term: due to the system's resistance to change, repeating the process every 5-10 years should suffice. In this way, more resources can be allocated to evaluative activity at the individual school level.

10. Evaluation should serve to improve instruction

Promoting the quality of instruction is a demanding task that involves conflicts. The completion of this task demands of education policy and education administration the implementation of differentiated measures. Evaluative measures must be organized in such a way that the desirable impulses and insights are not counteracted by undesirable side effects. For this reason, they are to be oriented to specific needs for information and action – within concrete fields of action. The main goal is to develop the everyday work of the school or of individual teachers in such a way that, e.g., by means of a better understanding of learning processes and difficulties, they can make more specific demands on and offer more specific support to students engaged in developing their own individual abilities. But, above all, the capacity and competence required for (self-)evaluation must be installed in the system itself on a long-term basis. Schools themselves must become 'learning systems'. To this end, teachers require competencies and methods appropriate for everyday application. In addition, they are dependent on the challenge presented by external perspectives and on the support from external competencies and resources. Here, it has proved useful for teaching staff to invite observers as 'critical friends' in order to confront internal perception with an external view. At the level of materials, assignment examples with commentaries and data from representative samples providing orientation are more useful than standardized comparative assignments for all classes that cannot relate to the specific situation on location (prerequisites of students, resources of the school).

In the final analysis, all such measures must contribute to further developing schools and concrete improvements of instruction, i.e., to supporting the teachers so that they can support the learners (Heide Bambach). For this reason, both efforts at evaluation and financial resources should be focused on studying and documenting models and prerequisites of successful instruction.

In particular, case studies of 'unexpectedly successful' persons or institutions can further the perception of conditions expedient to learning and to stimulate and support analogous developments at other places. Such studies are important at two levels:

- case studies of students who develop successfully at school despite unfavorable preconditions (individual handicaps, milieu remote to education, social disadvantages, kept back at the beginning of their schooling, repetition of a class);

- case studies of schools that work successfully despite difficult circumstances (substantial socio-economic problems in the area served by the school, large proportion of children with a different native language, lack of young teachers).

Thus, the priority should be on measures aiming at developing quality. A purely descriptive stocktaking of well-known problems (disadvantages of migrants, children from lower social strata) results in short-term public attention, but in little else. A higher priority than isolated comparisons of effects must be the analysis of processes and conditions. Only such analysis can identify reasons behind problems and substantiate suggestions for intervention.

Admittedly, it is important to observe the development of basic achievement in major school subjects. But it can have unfortunate consequences if this perspective is so predominant and so reductive in public opinion that other major aspects of the quality of schooling are ignored. Yet, this is precisely the situation at the moment. Output-oriented tests seem more appealing to education policy than process-oriented quality control measures because they are relatively inexpensive, can be prescribed from external agencies, can be put into practice rapidly, and produce highly visible results.[68]

References

AERA (2003) Position Statement on HIGH STAKES TESTING In PreK-12 Education. http://www.eval.org/hst3.htm [Abruf 9.10.2003].

Aldrich, R. (2000): Educational standards in historical perspective. In: Goldstein/ Heath (2000, 39-67).

Amrein, A. L./ Berliner, D. C. (2002): High-stakes testing, uncertainty, and student learning. In: Education Policy Analysis Archives, Vol. 10, No. 18. [http://epaa.asu.edu/epaa/v10n18/].

Aldrich, R. (2000): Educational standards in historical perspective. In: Goldstein/ Heath (2000, 39-67).

Amrein, A. L./ Berliner, D. C. (2002): High-stakes testing, uncertainty, and student learning. In: Education Policy Analysis Archives, Vol. 10, No. 18. [http://epaa.asu.edu/epaa/v10n18/].

Ascher, C. (1996): Performance contracting: A forgotten experiment in school privatization. In: Phi Delta Kappan, Vol. 77, No. 9, 615-621.

Bartnitzky, H., u. a. (1999): Zur Qualität der Leistung – 5 Thesen zu Evaluation und Rechenschaft der Grundschularbeit. Grundschulverband -- Arbeitskreis Grundschule e. V.: Frankfurt. Auch in: Schmitt (1999, 165-196).

Baumert, J., u. a. (Hrsg.) (2000b): TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie – Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn. Bd. 1: Mathematische und naturwissenschaftliche Grundbildung am Ende der gymnasialen Oberstufe. Leske+Budrich: Opladen.

Baumert, J., u. a. (Hrsg.) (2001): PISA 2000 – Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich. Leske + Budrich: Opladen.

Baumert, J., u. a. (2003): PISA 2000 – Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland. Leske+Budrich: Opladen.

Becker, D. H. (2002): Empirische Befunde zum Kerncurriculum: Implementierungsbedingungen und Effekte. In: Böttcher/Kalb (2002, 84-122).

Böttcher, W. (2003): Bildung, Standards, Kerncurricula. Ein Versuch, einige Missverständnisse auszuräumen. In: Die Deutsche Schule, 95, 2003, 2, 168-171.

Böttcher, W./ Hirsch, E. D. (1999): Über die Notwendigkeit eines verbindlichen Kerncurriculums. In: Die Deutsche Schule, 91. Jg., H. 3, 299-310.

Böttcher, W./ Kalb, P. E. (Hrsg.) (2002): Kerncurriculum – Was Kinder in der Grundschule lernen sollen. Beltz Pädagogik: Weinheim/ Basel.

Bos, W., u. a. (Hrsg.) (2003): Erste Ergebnisse aus IGLU. Schülerleistungen am Ende der vierten Jahrgangsstufe im internationalen Vergleich. Waxmann: Münster u. a.

[68] According to an argument proposed by Linn (2000, 4), which seems quite plausible in the light of the hectic responses to PISA.

Bos, W., u. a. (Hrsg.) (2004): Einige Länder der Bundesrepublik Deutschland im nationalen und internationalen Vergleich. Waxmann: Münster.

Brinkmann, E. (1997): Rechtschreibgeschichten – Zur Entwicklung einzelner Wörter und orthographischer Muster über die Grundschulzeit hinweg. Bericht No. 35 des Projekts OASE, FB 2 der Universität: Siegen.

Brinkmann, E. (2003) : „Farrat da war nichz Schwirich...“: In: Brinkmann u. a. (2003, 147-154).
Brinkmann, E., u. a. (Hrsg.) (2003): Kinder schreiben und lesen. Beobachten – Verstehen – Lehren. DGLS-Jahrbuch „Lesen und Schreiben“ Bd. 10. Deutsche Gesellschaft für Lesen und Schreiben. Fillibach-Verlag: Freiburg.

Brügelmann, H. (1978): Veränderungen des Curriculum auf seinem Weg vom Autor zum Kind. In: Zeitschrift für Pädagogik, 24. Jg., H. 4, 601-618.

Brügelmann, H. (1999): Was leisten unsere Schulen? Qualität und Evaluation von Unterricht in der Diskussion. Kallmeyersche Verlagsbuchhandlung: Seelze.

Brügelmann, H. (2003a): Das kurze Gedächtnis großer Reformer. Anmerkungen zum Beitrag von Wolfgang Böttcher in diesem Heft. In: Die Deutsche Schule, 95. Jg., H. 2, 168-171.

Brügelmann, H. (2003b): In fünf Jahren... Kerncurricula, Bildungsstandards und Leistungstests. in: Neue Sammlung, 43. Jg., H. 2, 235-237.

Brügelmann, H. (2003c): Grundlegende Leseleistungen und der „Karawanen-Effekt“ in der Grundschule. Zentrale Befunde aus dem Projekt LUST an der Universität Siegen. In: Grundschulverband aktuell Nr. 84 (November 2003, 19-25). s. a. www.uni-siegen.de/~agprim/lust

Brügelmann, H. (2004g): Lese-/ Schreibförderung nach PISA, IGLU und LUST: Was heißt eigentlich 'funktional alfabetisiert'? Ms. für Alfa-Forum, 18. Jg., Nr. 54 (Sommer 2004).

Brügelmann, H. (2004/05): Moden, Mythen und Modelle in der Pädagogik. Probleme von Erziehung und Unterricht aus der Sicht der Forschung. Libelle: CH-Lengwil (i.V.).

Brügelmann, H./ Richter, S. (Hrsg.) (1994): Wie wir recht schreiben lernen. Zehn Jahre Kinder auf dem Weg zur Schrift. Libelle Verlag: CH-Lengwil (2. Aufl. 1996).

Brügelmann, H./ Balhorn, H./ Füssenich, I. (Hrsg.) (1995): Am Rande der Schrift. Zwischen Mehrsprachigkeit und Analfabetismus. DGLS-Jahrbuch Bd. 6. Libelle Verlag: CH-Lengwil.

Carlson, D. (1979): Statewide assessment in California. In: Studies in Educational Evaluation, Vol. 5, No. 1, 55-75.

CERI (1978): National statement - United Kingdom. Vervielf. Ms. (SME/ET/78.80).OECD: Paris

Clare, J. (2004): Primary league tables „failing pupils“. In: News Telegraph v. 5.2.2004

Clarke, M. et al. (2000): High Stakes Testing and High School Completion. In: The National Board on Educational Testing and Public policy, Vol. 1, No. 3 (January 2000). ➔ http://www.bc.edu/research/nbetpp/publications/v1n3.html [Abruf: 13.2.2004]

Darling-Hammond, L. (2003): Standards and Assessments: Where We Are and What We Need. ➔ http://www.tcrecord.org/Content.asp?ContentID=11109 [Abruf: 15.2.2004]

Deutscher Bildungsrat (1974): Zur Förderung praxisnaher Curriculumentwicklung. Empfehlungen der Bildungskommission. Bundesdruckerei: Bonn.

Die Grünen – Bündnis 90 (Hrsg.) (2003): Bessere Schulen durch Bildungsstandards? Dokumentation einer öffentlichen Anhörung im Landtag von Baden-Württemberg am 21.3.2003.

Figlio, D./ Winicki, J. (2003): Food for thought: The effects of school accountability plans on school nutrition. NBER Working Paper No. 9319.

Gallin, P./ Ruf, U. (1998): Sprache und Mathematik in der Schule. Auf eigenen Wegen zur Fachkompetenz. Illustriert mit sechzehn Szenen aus der Biographie von Lernenden. Kallmeyer: Seelze (1. Aufl. Zürich 1990).

Gifford, B. G./ O’Connor, M. C. (eds.) (1992): Changing assessments: Alternative views of aptitude, achievement and instruction. Kluwer: Boston.

Götz, M. (1997): Die Grundschule in der Zeit des Nationalsozialismus. Klinkhardt Verlag: Bad Heilbrunn.

Goldstein, H./ Heath, A. (eds.) (2000): Educational standards. Proceedings of the British Academy. Vol. 102. Oxford University Press: Oxford.

Grigg, W. S., et al. (2003) : The nation's report card : Reading 2002. National Center for Educational Statistics. U.S. Department of Education: Washington, D. C.

Grundschulverband (2003): Bildungsansprüche von Grundschulkindern - Standards zeitgemäßer Grundschularbeit. In: Grundschulverband aktuell, Nr. 81, s. a. ➔ www.grundschulverband.de

Haag, L./ Stern, E. (2000): Non scholae sed vitae discimus? Auf der Suche nach globalen und spezifischen Transfereffekten des Lateinunterrichts. In: Zeitschrift für pädagogische Psychologie, 14. Jg., 146-157.

Hameyer, U./ Heckt, D. (2004): Bildungsstandards: „die" Lösung? In: Grundschule Special STANDARDS. Westermann: Braunschweig, 3-6.

Herrmann, U. (2004): Schule im Jahre IV nach PISA. Ein 10-Punkte-Programm gegen illusionäre „Bildungsstandards". Ms. für: Pädagogik, 56. Jg., H. 4.

Heymann, H. W. (2003): Why teach mathematics – A focus on general education. Kluwer: Dordrecht et al.

Hirsch, E. D. (1997): Cultural literacy. What every American needs to know. Boston.

House, E. R. (1975): Accountability in the U.S.A. In: Cambridge Journal of Education, Vol. 5, No. 2, 71-78.

Ingenkamp, K. (Hrsg.) (1971): Die Fragwürdigkeit der Zensurengebung. Beltz: Weinheim.

Jacobsen, S.(1989): Merit pay incentives in teaching. In: Weis et al. (1989, 111-128).

Kahl (2004): Vertrauen, Respekt, Selbstständigkeit. Die neuen finnischen Bildungsstandards sind auf Deutsch erschienen und begeistern immer mehr Schulen hierzulande. In: Süddeutsche Zeitung, Nr. 74 v. 29.3.2004, 10.

Kay, B. W. (1979): Processes of accountability in education -- England and Wales. Working Paper for O.E.C.D. Vervielf. Ms. (SME/ET/79.41). OECD: Paris.

Klieme, E., u. a. (2003): Zur Entwicklung nationaler Bildungsstandards. Eine Expertise. Deutsches Institut für Internationale Pädagogische Forschung: Frankfurt.

Köller, O. (2002): Des Schülers Leid, des Lehrers Freud. Schulnoten sind nötig und besser als ihr Ruf. In: Schule – Wissen – Bildung. Klett ThemenDienst Nr. 16: Dezember 2002, 7-10.

Kohn, A. (2002): The worst kind of cheating. In: Streamlined Seminar (National Association of Elementary School Principals), Vol. 21, No. 2 (Winter 2002/03.

Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee Value Added Assessment System. Educational Evaluation and Policy Analysis, Vol. 25, No. 3, 287-298.

Ladd, H.F./ Fiske, E.B. (2003). Does competition improve teaching and learning? Evidence from New Zealand. Educational Evaluation and Policy Analysis, Vol. 25, No. 1, 97-112.

Lehmann, R.H., u. a. (1995): Leseverständnis und Lesegewohnheiten deutscher Schülerinnen und Schüler. Beltz: Weinheim/ Basel.

Leßmann, B. (2003): Einsichten in die US-amerikanische Schulwirklichkeit. Aussichten für PISA-Konsequenzen in Deutschland? Wegweisung oder Warnung?! In: Grundschulunterricht, H. 9/2003, 45-49.

Linn, R. L. (2000): Assessments and acountability. In: Educational Researcher, Vol. 29, No. 2, 4-15.

MacDonald, B./ Walker, R. (1976): Changing the curriculum. Open Books: London.

Marshak, D. (2003): No child left behind : A foolish race into the past. In: Phi Delta Kappan International ➔ www.pdkintl.org/kappan/k0311ma2.htm

Martin, M. O., et al. (eds.) (2003) : PIRLS 2001. Technical report. Boston College: Chestnut Hill, MA.

Murnane, R./ Cohen, D. (1986): Merit pay and the evaluation problem: Why most merit pay plans fail and a few survive. In: Harvard Educational Review, Vol. 56, No. 1, 1-17.

NAEP (2001): The Nation's report card. Fourth grade reading highlights 2000. National Center for Education Statistics. U. S. Department of Education: Washington.

OECD & Statistics Canada (ed.) (1995b): Grundqualifikationen, Wirtschaft und Gesellschaft. Ergebnisse der ersten internationalen Untersuchung von Grundqualifikationen Erwachsener. Paris/ Ottawa (engl. 1995).

OECD (ed.) (2001): Knowledge and skills for life: First results from the OECD Programme for International Student Assessment (PISA). Organisation for Economic Co-Operation and Development : Paris.

OECD (2002): PISA 2000.Technical report. Organisation for Economic Co-Operation and Development: Paris.

O'Reilly, B. (2002): Why Edison doesn't work. In: Fortune, December 4, 2002 (www.fortune.com/fortune/articles/0,15114,395208,00.html).

Pedulla, J. J., et al. (2003): Percieved effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers. National Board on Educational Testing and Public Policy. Lynch School of Education: Boston College; s.a. www.bc.edu/research/nbetpp/reports.html [Zugriff 2.4.03]

Popham, W. J. (2001): The truth about testing: An educator's call to action. Association for Supervision and Curriculum Development: Alexandria.

Popham, W. J. (2002): Right task - wrong tool. Today's standardized tests are not the best way to evaluate schools or students. ➔ http://www.asbj.com/2002/02/0202coverstory.html

Rabenstein, R., u.a. (1989): Leistungsunterschiede im Anfangsunterricht. Heft 68. Institut für Grundschulforschung/ Universität: Nürnberg.

Ratzka, N. (2003): Mathematische Fähigkeiten und Fertigkeiten am Ende der Grundschulzeit – Empirische Studien im Anschluss an TIMSS (Phil. Diss. FB 2 der Universität Siegen). Franzbecker: Hildesheim/ Berlin.

Ratzki, A. (2003): Wer ist schuld, wenn Bildungsstandards nicht erfüllt werden? In: Lernende Schule, 6. Jg., H. 24, 30-31.

Rauin, U. (2003): Bildungsstandards in BW – Sachstand und Anforderungen. In: Die Grünen – Bündnis 90 (2003, 29-38).

Resnick, L. B./ Resnick, D. P. (1992): Assessing the thinking curriculum: New tools for educational reform. In: Gifford/ O'Connor (1992, 37-75).

Rhoades, K./ Madaus, G. (2003): Errors in standardized tests: A systemic problem. National Board on Educational Testing and Public Policy. Lynch School of Education: Boston; s.a. http://www.bc.edu/research/nbetpp/ (Zugriff 14.6.2003).

Richter, S. (1992): Die Rechtschreibentwicklung im Anfangsunterricht und Möglichkeiten der Vorhersage ihrer Störungen. Phil. Diss. FB 12 der Universität Bremen. Verlag Dr. Kovac: Hamburg.

Rossa, D./ Rossa, M. (1995): Erstunterricht in Neuseeland. In: Brügelmann u. a. (1995, 142-147).

Sacks, P. (1999): Standardized minds. The high price of America's testing culture and what we can do to change it. Cambridge, MA: Perseus Publishing.

Schiller, K.S. & Muller, C. (2003). Raising the bar and equity? Effects of state high school graduation requirements and accountability policies on students' mathematics course taking. In: Educational Evaluation and Policy Analysis, Vol. 25, No. 3, 299-318.

Schmitt, R. (Hrsg.) (1999): An der Schwelle zum dritten Jahrtausend. BundesGrundschulKongress 1999. Grundschulverband – Arbeitskreis Grundschule: Frankfurt.

Schrader, F.-W./ Helmke, A. (2001): Alltägliche Leistungsbeurteilung durch Lehrer. In: Weinert (2003, 43-58).

Schwanitz, D. (1999): Bildung – Alles, was man wissen muss. Eichborn-Verlag: Frankfurt a. M.

Schwarz, B./ Prange, K. (Hrsg.) (1997): Schlechte Lehrer/innen. Zu einem vernachlässigten Aspekt des Lehrberufs. Beltz: Weinheim/ Basel.

Schwarz, P. (2003): Veras Wahrheit. Vergleichsarbeiten in Klasse 4. In: Frankfurter Rundschau v. 4.6.2003.

Shepard, L. (1979): Purposes of Assessment. In: Studies in Educational Evaluation, Vol. 5, No. 1, 13-16.

Simon, J. (1979): What and who is APU? In: Forum, Vol. 22, No. 1, 7-11.

Stecher, B. M./ Barron, S. (2001): Unintended consequences of test-based accountability when testing in „milepost“ grades. In: Educational Assessment, Vol. 7, No. 4, 259-282.

Strunz, F. (2003): Kein Ende um Latein. Der totgesagte Park. In: Neue Sammlung, 43. Jg., H.4, 561-581.

Terhart, E. (1997): Gute Lehrer -- schlechte Lehrer. In: Schwarz / Prange (1997, 34-85).

Weinert, F. E. (Hrsg.) (2001): Leistungsmessungen in Schulen. Beltz/ Weinheim.

Weinert, F.E., u.a. (Hrsg.) (1974): Funk-Kolleg Pädagogische Psychologie. Bd. 1 und 2. Fischer Taschenbücher 6115/ 6116: Frankfurt.

Weis, L. et al. (eds.) (1989) : Crisis in teaching. Perspectives on current reforms. New York.

Wespel, M. (2004): Von der Leitidee „Sprach- und Schriftkultur“ zu Kompetenzen im Sprechen, Schreiben und Lesen. In: Grundschule Special STANDARDS. Westermann: Braunschweig, 7-10.

Westphalen, K. (2003): Latein oder Französisch? Überlegungen zum Bildungswert der zweiten Fremdsprache – Replik auf eine empirische Untersuchung. In: Forum Classicum, 46. Jg., 3-11.

Zielinski, W. (1974): Die Beurteilung von Schülerleistungen. In: Weinert u. a. (1974, 877-900).